Relational Information Moderates Approach-Avoidance Instruction Effects on Implicit Evaluation

Pieter Van Dessel[1]

Jan De Houwer[1]

Colin Tucker Smith[2]

[1]Ghent University, Belgium

[2]University of Florida, USA

Correspondence concerning this article should be addressed to Pieter Van Dessel, Ghent University, Department of Experimental-Clinical and Health Psychology, Henri Dunantlaan 2, B-9000 Ghent (Belgium). E-mail: Pieter.vanDessel@UGent.be.

Abstract

Previous research demonstrated that instructions to approach one stimulus and avoid another stimulus can result in a spontaneous or implicit preference for the former stimulus. In the current study, we tested whether the effect of approach-avoidance instructions on implicit evaluation depends on the relational information embedded in these instructions. Participants received instructions that they would move towards a certain non-existing word and move away from another non-existing word (self-agent instructions) or that one non-existing word would move towards them and the other non-existing word would move away from them (stimulus-agent instructions). Results showed that self-agent instructions produced stronger effects than stimulus-agent instructions on implicit evaluations of the non-existing words. These findings support the idea that propositional processes play an important role in effects of approach-avoidance instructions on implicit evaluation and in implicit evaluation in general.

Keywords: approach-avoidance, instructions, implicit evaluation, attitudes, propositional theory

Relational Information Moderates Approach-Avoidance Instruction Effects on Implicit Evaluation

As Zajonc (1980) argued in his seminal paper, people often evaluate stimuli in a spontaneous manner. Research has shown that such spontaneous or implicit evaluations are an important determinant of behavior (e.g., Greenwald, Poehlman, Uhlmann, & Banaji, 2009) and play a crucial role in a number of important psychological phenomena including psychopathology (Roefs et al., 2011), addiction (Wiers & Stacy, 2006), and social interaction (Fazio & Olson, 2003). Hence, understanding how implicit evaluations are acquired and activated is an important aim of psychological science. Cognitive theories of evaluation have traditionally assumed that implicit evaluations reflect the automatic activation of associations between representations in memory (for a review, see Hughes et al., 2011). Because associations are assumed to form automatically when two events co-occur, much research on the acquisition and change of implicit evaluations has employed paradigms in which stimuli are repeatedly paired with valenced stimuli (EC: Hofmann et al., 2010) or with valenced actions (approach-avoidance training: Kawakami, Phills, Steele, & Dovidio, 2007).

Recent studies, however, have established that changes in implicit stimulus evaluations can occur not only as the result of repeated pairings but also on the basis of mere instructions (De Houwer, 2006; Van Dessel, De Houwer, Gast, & Smith, 2015). For example, studies on the effects of approach-avoidance (AA) instructions have shown that participants who are instructed to approach certain stimuli and avoid other stimuli exhibit more positive implicit evaluations of to-be-approached stimuli than of to-be-avoided stimuli even if they never actually perform the AA actions. There is even evidence that these instruction-based effects on implicit evaluation can occur under certain conditions of automaticity. For instance, AA instructions influence implicit

evaluations even when participants do not consider the acquired information a valid basis for their evaluation (as indicated by the fact that they do not incorporate this information in their explicit evaluation; Van Dessel, De Houwer, Gast, Smith, & De Schryver, 2016).

Effects of AA instructions on implicit evaluation pose a challenge to a particular type of associative models that assume that (a) implicit evaluations reflect the automatic activation of associations in memory and (b) these associations are formed as the result of a slow-learning process that capitalizes on repeated co-occurrences (Rydell & McConnell, 2006; Smith & DeCoster, 2000). Yet, instruction-based AA effects are consistent with propositional models, which assume that propositions, rather than associations, guide implicit evaluation (e.g., De Houwer, 2009, 2014; Mitchell, De Houwer, & Lovibond, 2009). When participants are instructed to approach or avoid a stimulus, they might generate propositions about these stimulus-action relations, and these propositions can influence their implicit evaluations of the stimuli (Van Dessel et al., 2016). For example, changes in implicit evaluations may occur as the result of AA instructions when participants infer that to-be-approached stimuli are more positive than to-be-avoided stimuli (e.g., because they know that people typically approach good things and avoid bad things) and the automatic retrieval of this propositional information influences implicit evaluation.

Importantly, a propositional model of implicit evaluation not only predicts that implicit evaluations can form as the result of a single instruction, but also that these effects should depend on the relational information embedded in these instructions (De Houwer, 2014). Propositions store information not only about the strength of the relationship between concepts but also about the nature of the relation (e.g., ‘I approach Stimulus A’; see Shanks, 2007). If propositions mediate implicit evaluation, then changes in implicit evaluation (e.g., due to instructions) could

depend on the relational meaning of the acquired information (stored as propositions). Hence, instructions that contain the same concepts (e.g., 'approach' and 'Stimulus A') might produce dissimilar effects on implicit evaluations if those concepts are related in a different manner.

We tested this prediction by giving participants instructions that differed not in the pairing of concepts (e.g., the stimulus and the AA action word), but in the agency relation specified in the instructions (i.e., who performs the AA action). Half of the participants received typical AA instructions which stated that *the participant* would perform a specific AA action in relation to a specific stimulus (i.e., move towards or away from a non-existing word). The other half of the participants received instructions which stated that *the stimulus* (i.e., the non-existing word) would perform the AA action in relation to the participant. We refer to the former instructions as 'self-agent instructions' and to the latter as 'stimulus-agent instructions'. Immediately following these instructions, participants' implicit evaluations of the two stimuli were registered with an Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998)[1]. We examined whether the two types of instructions have a differential impact on implicit stimulus evaluations.

From the perspective of a propositional account, instruction effects on implicit evaluation should depend on the extent to which instructions allow for the acquisition of the propositional information that a specific stimulus is positive or negative. Participants may easily infer that the stimuli they approach are more positive than the stimuli they avoid because this is consistent with their previous learning history (i.e., most often, positive stimuli are approached and negative stimuli are avoided; see also Van Dessel et al., 2016). However, it is less certain that participants

[1] AA instruction effects have been observed on a number of implicit and explicit evaluation measures (see Van Dessel et al., 2015). The current study uses the IAT because this is currently the most widely used method to measure implicit evaluations. The IAT captures implicit evaluations in the sense that it registers evaluative responses under conditions that are typically associated with automatic processes (e.g., under time pressure, in the absence of evaluation goals, …: see De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009).

infer that an approaching stimulus is more positive than an avoiding stimulus because typically both pleasant and unpleasant stimuli can approach or avoid a person (see also Hsee, Tu, Lu, & Ruan, 2014). Hence, from the perspective of this propositional account, there are good reasons to predict that self-agent instructions should more strongly influence implicit evaluations than stimulus-agent instructions. As we will discuss in more detail later on, such a result would not only reveal an important moderator of AA instruction effects but would also have implications for theoretical accounts of those effects.

Method

Participants and Design

A total of 1306 English-speaking volunteers participated online via the Project Implicit research website (https://implicit.harvard.edu). We stopped the data-collection when at least 1000 participants had completed all measures of the experiment to ensure that we would have sufficient statistical power to detect even small effects after data-exclusion (power > .80 to detect an effect size of $d = 0.20$). All data were collected in one shot without intermittent data analysis. Overall dropout rate was 29.5%. The dropout rates were comparable across the two conditions: 30.8% in the self-agent condition and 28.1% in the stimulus-agent condition, $\chi^2(1) = 1.20$, $p = .27$. Hence, there was no evidence for condition-dependent attrition.

In line with the standard treatment of Project Implicit data (e.g., Smith, De Houwer, & Nosek, 2013), data-exclusion involved removing participants who (a) did not fully complete all questions and tasks (190 participants; i.e., 14.6%), (b) had error rates above 30% when considering all IAT blocks or above 40% for any one of the critical IAT test blocks (31 participants; i.e., 2.4%), or (c) responded faster than 400 ms on more than 10% of the IAT trials (29 participants; i.e., 2.2%). Analyses were performed on the data of 1056 participants (653

women, mean age = 33, SD = 14). Table 1 provides the number of included and excluded participants in each of the experimental conditions. The proportion of excluded participants did not differ significantly between conditions, $\chi^2(3) = 3.26$, $p = .35$. Note that including the data from all participants in the analyses did not result in any shift in significance for any of the reported effects. A full description of these results can be found at https://osf.io/d3tpj/. At this online repository we also provide a link to the online study as well as all data of the study and data analysis scripts.

Procedure

Upon being assigned to this study, participants were informed that they would participate in an experiment that would involve two meaningless words: UDIBNON and BAYRAM. Half of the participants then read the self-agent instructions:

> *You will perform a task in which you will move towards BAYRAM and you will move away from UDIBNON. It is very important to remember which action belongs to which word. You will need this information to complete the task successfully. Later on we will explain to you exactly how you will be able to perform this task. For now, it is crucial that you remember that you will move towards BAYRAM and move away from UDIBNON. Before we present these words and start the task, you will complete a categorization task. This will last about 5 minutes. Make sure that during that task you do not forget the instructions of the next task. Please press 'Continue' when you have memorized the instructions and are ready to begin the categorization task.*

The other half of the participants read the stimulus-agent instructions, which were identical with the exception of the two sentences that specified the agency relation. These sentences now indicated that participants would "*perform a task in which BAYRAM will move*

towards you and UDIBNON will move away from you". Participants were prompted to remember this information with the following sentence: *"For now, it is crucial that you remember that BAYRAM will move towards you and UDIBNON will move away from you".* Note that the assignment of the words to the approach or avoidance action was counterbalanced across participants and across instruction conditions.

The reaction time task that followed was an IAT in which participants categorized attribute words as 'positive' or 'negative' and target words UDIBNON and BAYRAM as 'Udibnon' or 'Bayram'. To avoid that the target stimuli were classified only on the basis of simple perceptual features, the words were presented in eight different combinations of font types (Arial Black and Fixedsys), capitalizations (uppercase and lowercase), and size (16pt and 18pt), resulting in 8 different stimuli (also see Zanon, De Houwer, Gast, & Smith, 2014). The IAT consisted of three practice blocks and two experimental blocks. Participants began the IAT with 20 practice trials sorting the target words and 20 practice trials sorting positive and negative stimuli. Next, participants completed 56 trials in which UDIBNON and positive stimuli shared a single response key and BAYRAM and negative stimuli shared a single response key (half of the participants completed the IAT in this way, while the other participants began by sorting BAYRAM and positive with the same key). Participants then practiced sorting target words with the response key assignment reversed for 40 trials and finally participants completed a second set of 56 trials in which UDIBNON shared a response key with negative and BAYRAM shared a response key with positive (or vice versa). If the participant made an error in categorizing, a red "X" appeared on the screen and the participant corrected their mistake in order to continue. Latencies were recorded until a correct response was made.

After the implicit evaluation task, participants rated their liking of each of the words by answering two questions for each word: “To what extent do you like BAYRAM/UDIBNON?” and “To what extent do you have warm feelings for BAYRAM/UDIBNON?”. Participants gave their ratings on 9-point Likert scales ranging from 1 (not at all warm; like not at all) to 9 (completely warm; like completely). The ratings were aggregated into a single score for each word by averaging the respective scores (mean Cronbach’s Alpha = .80, *SD* = 0.01).

Next, participants were asked to complete a manipulation check for each word. Participants who had received self-agent instructions were asked what they would have to do according to the instructions when seeing the word UDIBNON or the word BAYRAM. Participants answered by selecting one of three options of a dropdown menu with “move towards the word”, “move away from the word” and “I don’t remember” as possible answers. Participants who had received stimulus-agent instructions were asked what would happen according to the instructions when seeing the word UDIBNON or the word BAYRAM. They answered by selecting one of three options: “the word would move towards me”, “the word would move away from me” or “I don’t remember”.

After completion of the manipulation check, participants were asked to report to what extent they had tried to form a mental image of the task in order to help them remember the instructions of this task. Participants answered by selecting a number between 0 (not at all) and 5 (to a great extent). Participants then indicated whether they were familiar with the words UDIBNON or BAYRAM (because they had previously participated in experiments that used these words as stimuli or because they were familiar with these words as instances of Turkish

language) by selecting 'yes' or 'no' from a dropdown menu.[2] Finally, participants were informed that it was not necessary to complete the previously instructed task and they were thanked for their participation.

Results

IAT scores

IAT scores were calculated using the D2-algorithm (Greenwald, Nosek, & Banaji, 2003), such that higher scores indicate a stronger preference for BAYRAM over UDIBNON. The Spearman-Brown corrected split-half reliability of the evaluative IAT score, calculated on the basis of an odd-even split, was $r(1054) = .85$. Across groups, participants displayed a small implicit preference for BAYRAM over UDIBNON ($M = 0.05$, $SD = 0.48$), $t(1055) = 3.65$, $p < .001$, $d = 0.11$. We performed a 2 (Type of AA Instructions: self-agent instructions, stimulus-agent instructions) x 2 (Content of AA Instructions: approach BAYRAM and avoid UDIBNON, approach UDIBNON and avoid BAYRAM) x 2 (Instruction Memory: no errors on questions that probed memory for the AA instructions: 77.2%, at least one error: 22.8%) analysis of variance (ANOVA) on the IAT scores. To account for the unbalanced design, we used type III sums of squares in this and all subsequent statistical analyses. The ANOVA revealed a main effect of Instruction Memory, $F(1,1048) = 13.19$, $p < .001$,$\eta^2 = 0.010$, indicating that participants preferred BAYRAM more when they had correct memory of the AA instructions, and a significant interaction of Instruction Memory and Content of AA Instructions, $F(1,1048) = 28.00$, $p < .001$, $\eta^2 = 0.022$. Importantly, we also observed a significant three-way interaction effect of Instruction

[2] Excluding the data from the 36 participants (i.e., 3.4% of the entire sample) who indicated that they were familiar with the words had no significant influence on any of the analyses reported and thus their data were retained.

Memory, Content of AA Instructions, and Type of AA Instructions, $F(1,1048) = 4.29$, $p = .038$, $\eta^2 = 0.004$. No other main or interaction effects were observed, Fs < 3.07, ps >.080, $\eta^2_{s} < 0.001$.

To examine the three-way interaction, we performed separate t-tests for participants with correct and incorrect memory of the instructions in the self-agent and stimulus-agent instruction conditions. In line with previous results (Van Dessel et al., 2015), participants who had received self-agent instructions and who had correct memory of these instructions preferred BAYRAM more when they had received instructions to move towards BAYRAM and move away from UDIBNON ($M = 0.32$, $SD = 0.44$) than when they had received instructions to move away from BAYRAM and move towards UDIBNON ($M = -0.23$, $SD = 0.46$), $t(393) = 12.03$, $p < .001$, $d = 1.21$, 95% confidence interval of the difference (CI diff) [0.46, 0.64]. Participants who had received stimulus-agent instructions and who had correct memory of the instructions preferred BAYRAM more when they had received instructions that BAYRAM would move towards them and UDIBNON would move away from them ($M = 0.18$, $SD = 0.45$) than when they had received instructions that BAYRAM would move away from them and UDIBNON would move towards them ($M = -0.12$, $SD = 0.44$), $t(418) = 6.84$, $p < .001$, $d = 0.67$, 95% CI diff [0.21, 0.38]. As expected on the basis of the propositional account, the effect of self-agent instructions was significantly stronger than the effect of stimulus-agent instructions, $F(1,811) = 15.92$, $p < .001$, $d_{diff} = 0.54$. According with previous studies (e.g., Van Dessel et al.), there were no significant effects for participants who did not correctly remember the instructions, ts < 0.56, ps > .57, ds < 0.10.

We also performed a linear regression analysis to investigate main and interaction effects of the two between-subjects categorical variables and the continuous predictor variable of Mental Simulation (i.e., participants' response to the mental simulation question: $M = 2.48$, $SD = 1.71$).

The three-way interaction effect of Content of AA Instructions, Type of AA Instructions, and Instruction Memory remained significant, $F(1,1037) = 4.18$, $p = .041$, $\eta^2 = 0.002$, and we did not observe any main or interaction effects involving the Mental Simulation variable, Fs < 3.41, ps > .064, η^2s < 0.001. Moreover, an ANOVA on the data of 164 participants with correct instruction memory who had indicated that they had not formed a mental image of the task (response on the mental simulation question = 0) still produced a main effect of Content of AA Instructions, $F(1,160) = 23.65$, $p < .001$, $\eta^2 = 0.129$. In the analysis of this strongly reduced data set, the interaction only showed a weak trend towards significance, $F(1,160) = 2.11$, $p = .149$, $\eta^2 = 0.011$. Both self-agent and stimulus-agent instructions caused significant changes in implicit evaluations (self-agent instructions: $t(78) = 4.10$, $p < .001$, $d = 0.92$, 95% CI [0.24, 0.70]; stimulus-agent instructions: $t(82) = 2.70$, $p = .009$, $d = 0.59$, 95% CI [0.07, 0.44]).[3]

Process Dissociation

To examine whether (self-agent and/or stimulus-agent) instructions can influence automatic components of IAT performance we also decomposed IAT performance into controlled and automatic components using Process Dissociation Procedures (PDP; Jacoby, 1991). This question is important because IAT scores are known to be sensitive to controlled processes and might therefore not always provide a pure index of implicit evaluations (De Houwer, Beckers, & Moors, 2007). It has been argued that PDP can be used to separate controlled and automatic components of IAT performance (Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005).[4]

[3] Analyses on participants' explicit rating scores revealed a similar pattern as obtained for implicit evaluations. Similar to previous studies (e.g., Van Dessel et al., 2016), self-agent instructions produced larger effects on implicit than on explicit evaluations and changes in implicit evaluations were not fully mediated by changes in explicit evaluations. Because our main research questions involved effects on implicit evaluations we report the results of the analyses for explicit rating scores in the Appendix.

[4] To gain information about the automaticity of the IAT effects we also tried to fit two multinomial processing tree (MPT) models that were recently designed to disentangle processes underlying IAT performance (i.e., the Quad

While realizing that also PDP have limitations, we opted to employ PDP in the current research in order to provide an additional, more stringent test of the idea that (certain types of) AA instructions can influence implicit evaluations(i.e., in addition to implicit *measures*).

In accordance with the procedures outlined by Payne (2001), we calculated automatic parameter estimates for IAT performance on the basis of participants' errors in the IAT (average error rate = 6.7%, *SD* = 5.1%). The automatic parameter estimate provides an indication of the extent to which participants' performance indicates that participants automatically evaluate BAYRAM as positive and UDIBNON as negative or vice versa. Estimates above 0.50 indicate that participants automatically respond more in accordance with the former knowledge. An ANOVA on automatic IAT parameters revealed a main effect of Content of AA Instructions, $F(1,1048) = 19.82$, $p < .001$, $\eta^2 = 0.001$, and an interaction effect of Content of AA Instructions and Type of AA Instructions, $F(1,1048) = 4.23$, $p = .040$, $\eta^2 = 0.001$. The three-way interaction effect of Content of AA Instructions, Type of AA Instructions, and Instruction Memory was marginally significant, $F(1,1048) = 3.73$, $p = .054$, $\eta^2 = 0.003$. For participants with correct memory and self-agent instructions, the automatic parameter estimates were significantly larger if they had received instructions to move towards BAYRAM and move away from UDIBNON ($M = 0.51$, $SD = 0.01$) than if they had received instructions to move away from BAYRAM and move towards UDIBNON ($M = 0.50$, $SD = 0.01$), $t(393) = 5.13$, $p < .001$, $d = 0.52$, 95% CI diff [0.004, 0.010]. For participants with correct memory and stimulus-agent instructions, we observed only a marginally significant difference in automatic parameter estimates between

---

Model, Conrey et al., 2015, and the ReAL model, Meissner & Rothermund, 2013). These models did not fit our data well. In the PDP model an automatic parameter of IAT performance is calculated only on the basis of the difference in the number of errors in compatible and incompatible IAT blocks. Though these procedures are less complex than the procedures of MPT models of IAT performance, it is consistent with these models in that they also assume that this difference in errors is strongly related to the more automatic processes that are involved in IAT performance.

participants with opposite instructions (BAYRAM moves towards you and UDIBNON moves away from you: $M = 0.50$, $SD = 0.01$; BAYRAM moves away from you and UDIBNON moves towards you: $M = 0.50$, $SD = 0.01$), $t(418) = 1.76$, $p = .079$, $d = 0.17$, 95% CI diff [-0.003, 0.006][5]. There were no significant effects for participants with incorrect memory of the instructions, ts < 0.69, ps > .49, ds < 0.12.

General Discussion

The current experiment was designed to test whether the effects of AA instruction on implicit evaluations depend on the relational information specified in the instructions. Results showed that instructions to move towards or away from certain non-existing words (i.e., self-agent instructions) caused bigger changes in implicit evaluations of these words than instructions in which the words were said to move towards or away from the participant (i.e., stimulus-agent instructions). PDP analyses indicated that only the former type of instructions had a significant effect on the automatic parameter of IAT performance, suggesting that only these typical AA instructions caused genuine changes in implicit evaluations. These results provide the first evidence that instructions that contain the same concepts produce different effects on implicit evaluations if those concepts are related in a different manner.

Our results are consistent with, and predicted by, a propositional account of AA instruction effects. From the perspective of a propositional account, participants may infer on the basis of AA instructions that to-be-approached stimuli are more positive than to-be-avoided

[5] To examine the robustness of this (marginally significant) effect we also performed Bayesian analyses (according to the procedures outlined by Rouder, Speckman, Sun, Morey, & Iverson, 2009). These analyses provide a Bayes Factor that gives an indication of how strongly the data support either the null hypothesis (BF_0; reflecting the absence of a significant effect) or the alternative hypothesis (BF_1; reflecting the presence of a significant effect). The Bayes Factor indicated anecdotal evidence for the null hypothesis, $BF_0 = 2.07$. In contrast, the Bayes Factor indicated strong evidence for the alternative hypothesis in the context of self-agent instructions, $BF_1 > 10000$.

stimuli. Because such inferences are less certain when instructions specify that it are the stimuli that approach or avoid the participant, self-agent instructions should produce bigger effects on implicit evaluations than stimulus-agent instructions. The observation that stimulus-agent instructions also had a significant effect on implicit evaluations, albeit to a lesser extent than self-agent instructions, was not predicted a priori by the propositional account. On the one hand, this result has to be interpreted with some caution because we did not observe a significant impact of stimulus-agent instructions on the automatic component of IAT performance as indexed using PDP. This suggests that the effects of stimulus-agent instructions might more strongly depend on controlled, non-automatic processes that involve the intentional use of the acquired information for evaluation (e.g., as the result of demand compliance). On the other hand, the mere presence of such an effect could be explained by the propositional account in a post-hoc fashion. One possible post-hoc explanation for this result is that people believe to some extent that stimuli that move towards them are typically more positive than stimuli that move away from them. Such a belief might be based on certain concrete experiences (e.g., receiving positive things such as gifts or money and removing negative things such as garbage or dirt). This post-hoc account could be tested in future research by explicitly asking participants to report their beliefs on this issue or by manipulating the extent to which participants are likely to believe that approaching stimuli are more positive than stimuli that move away from them.

Whereas the moderating effect of relational information on AA instructions effects was predicted on the basis of propositional models of implicit evaluation, it raises questions about specific types of associative models. First, our results do not fit well with associative accounts of these effects which assume that (automatic) changes in implicit evaluations are due to the mere pairing of stimuli and action words in the instructions (e.g., Field, 2006). Though this account

could explain why both stimulus-agent and self-agent instructions influence implicit evaluation, it cannot explain the stronger effect of stimulus-agent instructions because both types of instructions contain the same pairing of stimuli and action words. For our findings to be reconciled with these accounts it seems necessary to assume that different action representations can be activated depending on the order of the words in the instructions (e.g., 'will move towards you' and 'you will move towards'). If these action representations differ in how strongly they are linked with valenced representations, this could allow for differences in associative transfer of valence to the non-word stimuli. Note, however, that this mechanism would require the operation of (propositional) processes that can integrate information about the order of elements (i.e., words). Second, our results also do not support associative accounts which assume that mental simulation of (repeated) AA actions in response to the stimuli determines AA instruction effects on implicit evaluation. AA instruction effects were not related to self-reports about the extent to which participants mentally simulated the AA action. Note, however, that it is still possible that mental simulation contributed to the effects and participants were simply not very good at reporting this mental simulation.

It is also possible that propositional and associative processes jointly contribute to AA instruction effects on implicit evaluation. As noted by a reviewer, a first possible explanation is that participants extract the relational content of the instructions on the basis of propositional processes. This could lead to the activation of a mental representation of the to-be-performed action that incorporates agency information (e.g., different nodes of "I approach it" and "it approaches me"). Representations of self-agent AA actions might have stronger mental links with representations of positive or negative valence than representations of stimulus-agent AA actions. This could in part be due to the fact that only self-agent representations are linked to

representations of the self, which are typically positive (Gawronski, Bodenhausen, & Becker, 2007). As a result, there might be more associative transfer of positive valence to stimulus representations when participants receive self-agent approach instructions than when they receive stimulus-agent instructions. It should be noted, however, that a recent study provided little evidence for the involvement of representations of the self in AA instruction effects (Van Dessel, Gawronski, Smith, & De Houwer, 2017). It is, however, also possible that self-agent action representations are more positive because they have stronger associations with other valenced representations (e.g., representations of motivational systems of approach and avoidance, Neumann, Förster, & Strack, 2003), which facilitates associative transfer of valence in the context of self-agent instructions. Future studies are required to test these explanations. As a final note with regard to this alternative explanation of our results, we would like to point out that self-agent and stimulus-agent action representations themselves are propositional in that they incorporate relational information (i.e., information about what element has the role of approaching). One could thus argue that this alternative account is also propositional in nature (De Houwer, 2014).

A second possible explanation is that associative processes are driven by propositional beliefs. Propositional processes might allow one to acquire the propositional belief that a stimulus is positive or negative and this could lead to the formation of mental associations that reflect this information. For instance, when one infers that a to-be-approached stimulus is more positive than a to-be-avoided stimulus, this might lead to the formation of an association between a representation of this stimulus and representations of positive or negative valence. The (automatic) activation of these associations might mediate IAT performance. Note that this alternative account assumes that both propositional and associative processes are involved in AA

instructions effects on implicit evaluations. For reasons of parsimony, a propositional account that can explain these results is to be preferred over an account that additionally postulates the existence of an entirely different second mechanism (i.e., association formation).

Regardless of the merits of the described explanations, we can conclude that our results put new and important constraints on any current and future explanation of AA instruction effects. They support the contribution of propositional processes to AA instruction effects on implicit evaluations and rule out specific associative explanations. Of course, our results cannot distinguish between the broad class of single-process propositional and the broad class of associative or dual-process models. Because these models have a high degree of flexibility, distinguishing between these models on the basis of a single set of data is difficult, if not impossible. However, the current results do allow us to further constrain evaluation models and to have greater confidence in predictions derived from these models. The continuing investigation of the effects of (AA) instructions on implicit evaluation may help to further elucidate the mechanisms involved in (implicit) evaluation.

Acknowledgments

Funding: Pieter Van Dessel is supported by a Postdoctoral fellowship of the Scientific Research Foundation, Flanders (FWO-Vlaanderen). Jan De Houwer is supported by Methusalem Grant BOF16/MET_V/002 of Ghent University and by the Interuniversity Attraction Poles Program initiated by the Belgian Science Policy Office (IUAPVII/33).

References

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173−1182. doi:10.1037/0022-3514.51.6.1173

Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. (2005). Separating multiple processes in implicit social cognition: The Quad-Model of implicit task performance. *Journal of Personality and Social Psychology*, *89*, 469-487.

De Houwer, J. (2006). Using the Implicit Association Test does not rule out an impact of conscious propositional knowledge on evaluative conditioning. *Learning and Motivation*, *37*, 176–187. doi:10.1016/j.lmot.2005.12.002

De Houwer, J. (2009). The propositional approach to associative learning as an alternative for association formation models. *Learning & Behavior, 37*, 1–20. doi: 10.3758/LB.37.1.1

De Houwer, J. (2014). A Propositional Model of Implicit Evaluation. *Social and Personality Psychology Compass, 8*, 342-353. doi:10.1111/spc3.12111

De Houwer, J., Beckers, T., & Moors, A. (2007). Novel attitudes can be faked on the Implicit Association Test. *Journal of Experimental Social Psychology, 43*, 972-978. doi: 10.1016/j.jesp.2006.10.007

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin, 135*, 347-368. doi:10.1037/a0014211

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition. research: their meaning and use. *Annual Review of Psychology*, *54*, 297–327. doi: 10.1146/ annurev.psych.54.101601.145225

Field, A. P. (2006). Is conditioning a useful framework for understanding the development and treatment of phobias? *Clinical Psychology Review, 26,* 857-875. doi:10.1016/j.cpr.2005.05.010

Gawronski, B., Bodenhausen, G. V., & Becker, A. P. (2007). I like it, because I like myself: Associative self-anchoring and post-decisional change of implicit evaluations. *Journal of Experimental Social Psychology*, *43*, 221-232. doi:10.1016/j.jesp.2006.04.001

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology*, *74*, 1464–1480. doi:10.1037/0022-3514.74.6.1464

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85*, 197–216. doi:10.1037/0022-3514.85.2.197

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97*, 17–41. doi:10.1037/a0015575

Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: a meta-analysis. *Psychological Bulletin, 136,* 390–421. doi:10.1037/a0018916

Hsee, C., Tu, Y., Lu, Z., & Ruan, B. (2014). Approaching aversion: Negative reactions toward approaching stimuli. *Journal of Personality and Social Psychology, 106*, 699–712. doi:10.1037/a0036332

Hughes, S., Barnes-Holmes, D., & De Houwer, J. (2011). the Dominance of Associative Theorizing in Implicit Attitude Research: Propositional and Behavioral Alternatives. *Psychological Record, 61*, 465–496.

Jacoby, L.L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30*, 513–541. doi:0.1016/0749596X(91)90025

Kawakami, K., Phills, C. E., Steele, J. R., & Dovidio, J. F. (2007). (Close) distance makes the heart grow fonder: Improving implicit racial evaluations and interracial interactions through approach behaviors. *Journal of Personality and Social Psychology, 92*, 957-971. doi:10.1037/0022-3514.92.6.957

Meissner, F., & Rothermund, K. (2013). Estimating the contributions of associations and recoding in the Implicit Association Test: The ReAL model for the IAT. Journal of Personality and Social Psychology, 104, 45–69. doi: 10.1037/a0030734

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning, *Behavioral and Brain Sciences, 32*, 183-198. doi: 10.1017/S0140525X09000855

Neumann, R., Förster, J., & Strack, F. (2003). Motor compatibility: The bidirectional link between behavior and evaluation. In J. Musch & K. C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion.* (pp. 371–391). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Payne, B.K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology, 81*, 181–192. doi:10.1037/00223514.81.2.181

Roefs, A., Huijding, J., Smulders, F. T. Y., MacLeod, C. M., de Jong, P. J., Wiers, R. W., & Jansen, A. T. M. (2011). Implicit measures of association in psychopathology research. *Psychological Bulletin, 137*, 149–193. doi: 10.1037/a0021729

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1-36. URL: http://www.jstatsoft.org/v48/i02/

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225–237. doi:10.3758/PBR.16.2.225

Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: a systems of reasoning analysis. *Journal of* Personality *and Social Psychology, 91*, 995–1008. doi:10.1037/0022-3514.91.6.995

Shanks, D. R. (2007). Associationism and cognition: human contingency learning at 25. *Quarterly Journal of Experimental Psychology, 60*, 291–309. doi: 10.1080/17470210601000581

Smith, E. R., & DeCoster, J. (2000). Dual-Process Models in Social and Cognitive Psychology: Conceptual Integration and Links to Underlying Memory Systems. *Personality and Social Psychology Review, 4*, 108–131. doi: 10.1207/S15327957PSPR0402_01

Smith, C. T., De Houwer, J., & Nosek, B. (2013). Consider the source: Persuasion of implicit evaluations is moderated by source credibility. *Personality and Social Psychology Bulletin,* 39, 193-205. doi:10.1177/0146167212472374

Van Dessel, P., De Houwer, J., Gast, A., & Smith, C. T. (2015). Instruction-Based Approach–Avoidance Effects: Changing Stimulus Evaluation via the Mere Instruction to Approach or Avoid Stimuli. *Experimental Psychology, 62,* 161-169. doi:10.1027/1618-3169/a000282

Van Dessel, P., De Houwer, J., Gast, A., Smith, C. T., & De Schryver, M. (2016). Instructing Implicit Processes: When Instructions to Approach or Avoid Influence Implicit but not Explicit Evaluation. *Journal of Experimental Social Psychology, 63,* 1-9. doi:10.1016/j.jesp.2015.11.002

Van Dessel, P., Gawronski, B., Smith, C. T., & De Houwer, J. (2017). Mechanisms underlying approach-avoidance instruction effects on implicit evaluation: Results of a preregistered adversarial collaboration. *Journal of Experimental Social Psychology, 69,* 23-32. doi:10.1016/j.jesp.2016.10.004

Wiers, R. W., & Stacy, A. W. (2006). Implicit cognition and addiction. *Current Directions in Psychological Science*, *15*, 292–296. doi: 10.1111/ j.1467-8721.2006.00455.x

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist, 35,* 151-175. doi:10.1037/0003-066X.35.2.151

Zanon, R., De Houwer, J., Gast, A., Smith, C. T. (2014). When does relational information influence evaluative conditioning? *Quarterly Journal of Experimental Psychology, 67,* 2105-2122. doi:10.1080/17470218.2014.907324

Table 1. *Total number of participants in each of the four experimental conditions as a function of meeting the inclusion criteria for data analyses.*

	self-agent instructions		stimulus-agent instructions	
	approach BAYRAM	approach UDIBNON	approach BAYRAM	approach UDIBNON
	N	N	N	N
Included	263 (20.1%)	230 (17.6%)	297 (22.7%)	266 (20.4%)
Excluded	69 (5.3%)	62 (4.8%)	58 (4.4%)	61 (4.7%)
	332 (25.4%)	292 (22.4%)	355 (27.2%)	327 (25.0%)

Appendix

Data-analyses on explicit rating scores

We observed a moderate correlation of explicit rating scores with evaluative IAT scores, $r(1054) = .35$, $p < .001$. Similar to the results of the IAT, explicit rating scores revealed a significant preference for BAYRAM over UDIBNON ($M = 0.67$, $SD = 2.22$), $t(1055) = 9.83$, $p < .001$, $d = 0.30$. The ANOVA on explicit rating scores revealed a main effect of Content of AA Instructions, $F(1,1048) = 48.11$, $p < .001$, $\eta^2 = 0.004$, an interaction of Content of AA instructions and Instruction Memory, $F(1,1048) = 30.20$, $p < .001$, $\eta^2 = 0.026$, and a three-way interaction of Content of AA instructions, Type of AA instructions and Instruction Memory $F(1,1048) = 5.60$, $p = .015$, $\eta^2 = 0.005$. Participants with correct instruction memory who had received self-agent instructions preferred BAYRAM more when they received instructions to approach BAYRAM and avoid UDIBNON ($M = 1.45$, $SD = 2.19$) than when they received instructions to avoid BAYRAM and approach UDIBNON ($M = -0.24$, $SD = 2.36$), $t(393) = 7.22$, $p < .001$, $d = 0.73$, 95% CI diff [1.21, 2.11]. Participants with correct instruction memory who had received stimulus-agent instructions preferred BAYRAM more when they received instructions that BAYRAM would approach them and UDIBNON would avoid them ($M = 1.13$, $SD = 2.03$) than when they received instructions that BAYRAM would avoid them and UDIBNON would approach them ($M = 0.14$, $SD = 2.18$), $t(418) = 4.85$, $p < .001$, $d = 0.47$, 95% CI diff [0.59, 1.40]. Importantly, similar to IAT scores, the effect of self-agent instructions on explicit ratings was significantly stronger than the effect of stimulus-agent instructions, $F(1,811) = 4.61$, $p = .032$, $\eta^2 = 0.005$.

An ANOVA that included Mental Simulation as covariate revealed a main effect of Content of AA Instructions, $F(1,1037) = 46.58$, $p < .001$, $\eta^2 = 0.063$, and an interaction of

Content of AA instructions and Instruction Memory, $F(1,1037) = 28.12$, $p < .001$, $\eta^2 = 0.006$. This interaction was qualified by an interaction of Content of AA instructions, Instruction Memory, and Mental Simulation, $F(1,1037) = 10.17$, $p < .001$, $\eta^2 = 0.009$. Participants with correct instruction memory exhibited a larger AA instruction effect when they had higher mental simulation scores, $F(1,804) = 4.87$, $p = .028$, $\eta^2 = 0.005$, whereas participants with incorrect instruction memory exhibited a smaller AA instruction effect when they had higher mental simulation scores, $F(1,233) = 7.10$, $p = .008$, $\eta^2 = 0.024$. Importantly, the three-way interaction of Content of AA instructions, Type of AA instructions and Instruction Memory remained significant, $F(1,1037) = 5.78$, $p = .016$, $\eta^2 = 0.004$, and we observed no other interaction effects with Mental Simulation, Fs < 0.97, ps > .32, η^2s < 0.001. Participants who indicated that they had not formed a mental image of the task still exhibited an AA instruction effect if they received self-agent instructions, $t(78) = 2.11$, $p = .038$, $d = 0.47$, 95% CI [0.06, 1.88], but not if they received stimulus-agent instructions, $t(82) = 0.85$, $p = .40$, $d = 0.19$, 95% CI [-0.47, 1.17]. The interaction of Content of AA instructions and Type of AA instructions, however, was not significant, $F(1,160) = 1.00$, $p = .319$, $\eta^2 = 0.006$.

Mediation analyses

To investigate the relationship between AA instruction effects on implicit and explicit evaluations, we performed mediation analyses with the lavaan package in R (version 0.5-16; Rosseel, 2012). We used the bootstrap method to estimate standard errors for the effects. We tested whether changes in explicit rating scores mediate the effect of self-agent and stimulus-agent AA instructions on IAT scores (see Figures A1, A2). Toward this end, IAT scores were simultaneously regressed on both AA instructions (approach BAYRAM and avoid UDIBNON, approach UDIBNON and avoid BAYRAM) and explicit rating scores (Baron & Kenny, 1986).

For participants in the self-agent instructions condition, the indirect effect of AA instructions on IAT scores with explicit rating scores as a mediator was statistically significant, β = .15, Z = 4.38, p <.001, 95% CI of β = [0.08,0.23], R^2_{ind} = 0.24. However, the effect of AA instructions on the IAT score remained statistically significant after controlling for changes in explicit evaluations, β = .72, Z = 9.24, p < .001, 95% CI of β = [0.55, 0.86], R^2_{dir} = 0.54. These results replicate previous findings that AA instruction effects on implicit evaluations are not fully mediated by changes in explicit evaluations (Van Dessel, De Houwer, Gast, & Smith, 2015; Van Dessel et al., 2016).

For participants in the stimulus-agent instructions condition, the indirect effect of AA instructions on IAT scores with explicit rating scores as a mediator was also statistically significant, β = .09, Z = 3.45, p <.001, 95% CI of β = [0.07,0.13], R^2_{ind} = 0.10. The effect of AA instructions on the IAT score remained statistically significant after controlling for changes in explicit evaluations, β = .12, Z = 3.12, p = .009, 95% CI of β = [0.03, 0.18], R^2_{dir} = 0.08.

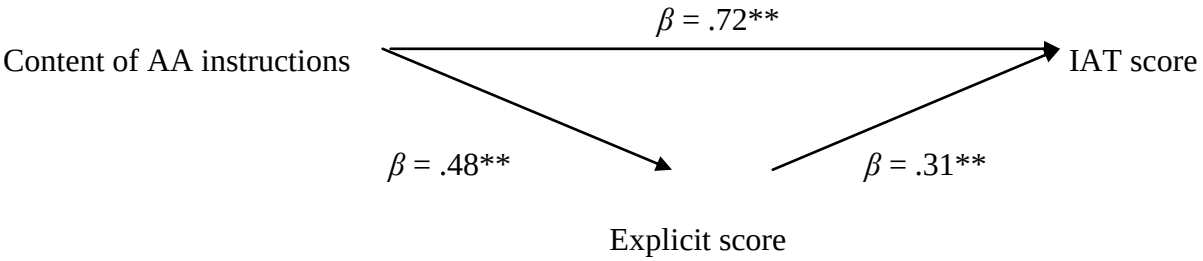

Figure A1. Estimates of mediation coefficients for participants who received self-agent instructions. * p < .05 ** p < .001.

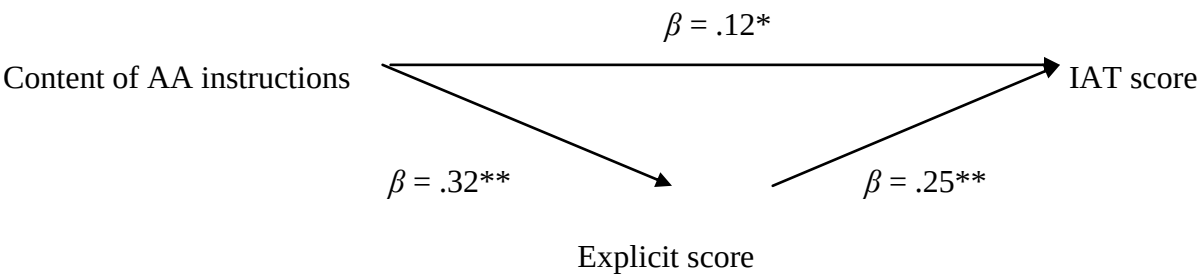

Figure A2. Estimates of mediation coefficients for participants who received stimulus-agent instructions. * $p < .05$ ** $p < .001$.